SOME MINIMAX INVARIANT PROCEDURES
FOR
ESTIMATING A CUMULATIVE DISTRIBUTION FUNCTION

By

Om P. Aggarwal

University of Washington

Technical Report No. 15

August 18, 1954

Contract N8onr-520 Task Order II
Project Number NR-042-038

Laboratory of Statistical Research
Department of Mathematics
University of Washington
Seattle, Washington

Some Minimax Invariant Procedures for Estimating a Cumulative Distribution Function

1. Summary. Some invariant procedures, which are essentially step functions, are considered as estimators of the cumulative distribution function of a univariate random variable on which a finite fixed number of observations are given, for various loss functions. Two principal classes of loss functions are considered and it is shown that for a special loss function in one class the optimum procedure is the usual sample cumulative function.

2. Introduction. Suppose that a sample X_1, X_2, ...X_n of a one-dimensional chance variable X is given. In a recent paper, Birnbaum [1] has discussed various techniques for deciding whether X has a completely specified continuous cumulative distribution function (c.d.f.), $H(x)=P(X\leq x)$. In this paper is discussed an allied problem, viz., that if $F(x)=P(X\leq x)$ is the unknown continuous c.d.f. of X and if $\hat{F}(x)_{X_1,\dots X_n}$ be an estimate of F(x) based on the sample $X_1,X_2,\dots X_n$, what would be the best estimate $\hat{F}$ when certain forms of the loss function are given.

Consider the loss function

$$(1) \qquad L(F,\hat{F})=\int_{-\infty}^{\infty}|F(x)-\hat{F}(x)|^r\,dx,$$

where r is an integer ≥ 1. It is almost obvious that the only invariant procedures for estimating F under the group of all one-to-one monotone transformations of the real numbers onto themselves which leave the sample values X_i $(i=1,2,\dots n)$ invariant are those which estimate F(x) by a step function

$$(2) \qquad \hat{F}(x)=\text{constant, say } c_j \quad \text{for } X^{(j)}\leq x<X^{(j+1)}$$

where $X^{(1)}<X^{(2)}<\dots<X^{(n)}$ are the ordered observations and $X^{(0)}$ and $X^{(n+1)}$ denote $-\infty$ and $+\infty$ respectively.

Using this estimate $\hat{F}$, we get

$$L(F,\hat{F}) = \sum_{j=0}^{n} \int_{X^{(j)}}^{X^{(j+1)}} \left|F(x)-c_j\right|^r dF(x)$$

(3)

$$= \frac{1}{r+1} \sum_{j=0}^{n} \left[(F(X^{(j+1)})-c_j)\left|F(X^{(j+1)})-c_j\right|^r - (F(X^{(j)})-c_j)\left|F(X^{(j)})-c_j\right|^r \right]$$

and the right hand side of this equation is a symmetric function of $F(X_1)$, $F(X_2)$, ..., $F(X_n)$ where $X_1, X_2, ..., X_n$ is the unordered sample. It follows now from a theorem of Birnbaum and Rubin [2] that $L(F,\hat{F})$ is a distribution free statistic and hence the risk R, being the expectation of L with respect to the distribution F, is constant and independent of F itself. We can thus take F to be a rectangular distribution over (0,1) and write

(4) $$R = E \sum_{j=0}^{n} \int_{X_j}^{X_{j+1}} \left|x-c_j\right|^r dx$$

where $X_1 < X_2 < ---- < X_n$ is an ordered sample of size n from this rectangular distribution over (0,1), X_0 and X_{n+1} denote 0 and 1 respectively, and the symbol E denotes that the expectation is taken with respect to the rectangular distribution over (0,1). In the rest of this paper, we shall use consistently the letter E to denote the fact that the expectation is to be taken with respect to the rectangular distribution over (0,1).

The same argument applies when the loss function is of the form

(5) $$L(F,\hat{F}) = \int_{-\infty}^{\infty} \frac{\left|F(x)-\hat{F}(x)\right|^r}{F(x)\left[1-F(x)\right]} dF(x)$$

and in this case by taking $\hat{F}$ as in (2) we obtain

$$(6)\qquad R = E\sum_{j=0}^{n}\int_{X_j}^{X_{j+1}} \frac{|x-c_j|^r}{x(1-x)}\,dx$$

where X_j; $j=0,1,\ldots n+1$ are the same as in (4).

It is obvious that since risk R is constant, a minimax procedure will be to choose c_j; $j=0,1\ldots n$, such that R is minimum. We consider in this paper the values of c_j when the loss function is of the form (1) for all integers $r \geqslant 1$ and when the loss function is of the form (5) for $r = 1$ and when r is an even integer $\geqslant 2$. The case when r is odd in (5) seems to be rather complicated.

3. The loss function $L(F,\hat{F}) = \int_{-\infty}^{\infty}\left[F(x) - \hat{F}(x)\right]^2 dF(x)$.

In this case

$$R = E\sum_{j=0}^{n}\int_{X_j}^{X_{j+1}} (x - c_j)^2 dx$$

$$(7)\qquad = 1/3 \sum_{j=0}^{n} E\left[(X_{j+1}^3 - X_j^3) - 3c_j(X_{j+1}^2 - X_j^2) \quad + 3c_j^2(X_{j+1} - X_j)\right]$$

Since the distribution of the j-th order statistic X_j in a sample of size n from the rectangular distribution over (0,1) is a Beta distribution with probability density

$$(8)\qquad p(y) = \frac{1}{B(j,n-j+1)}\quad y^{j-1}(1-y)^{n-j}\,,\ 0 \leq y \leq 1\,,$$

it is easily seen that for any positive integer r,

$$(9)\qquad E(X_j^r) = \frac{j(j+1)\ldots(j+r-1)}{(n+1)(n+2)---(n+r)}\,,$$

and

$$(10)\qquad E(X^r_{j+1}-X^r_j) = \frac{r(j+1)(j+2)\ldots(j+r-1)}{(n+1)(n+2)\ldots(n+r)}\ ,\ r\neq 1$$

$$\frac{1}{n+1}\ \text{for } r\ \ 1.$$

It will be useful to remark that (10) holds for all j; $j = 0, 1, \ldots, n$.

Substituting in R we obtain after some simplification,

$$R = 1/3 - \frac{2}{(n+1)(n+2)}\sum_{j=0}^{n}(j+1)c_j + \frac{1}{n+1}\sum_{j=0}^{n} c_j^{\,2}.$$

$$(11)\qquad = \frac{1}{6(n+2)} + \frac{1}{n+1}\sum_{j=0}^{n}\left(c_j - \frac{j+1}{n+2}\right)^2.$$

We see thus that R is minimized by choosing

$$(12)\qquad c_j = \frac{j+1}{n+2}\quad ;\qquad j = 0,1,\ldots n.$$

and hence the minimax invariant procedure is to estimate F(x) by

$$(13)\qquad \hat{F}(x) = \frac{j+1}{n+2}\quad ;\quad X_j \leq x < X_{j+1}\ ,\ j=0,1,\ldots,n,$$

where $(X_1, X_2, \ldots X_n)$ is the ordered sample and X_0 and X_{n+1} stand for $-\infty$ and $+\infty$ respectively.

The minimum risk corresponding to this procedure is seen to be $1/6(n+2)$. It is of some interest to note that the risk corresponding to the usual procedure of taking $c_j = j/n$ is given by $1/6n$.

4. The loss function $L(F,\hat{F}) = \int_{-\infty}^{\infty} |F(x) - \hat{F}(x)|\, dF(x)$.

For this case

$$(14)\qquad R = E\sum_{j=0}^{n}\int_{X_j}^{X_{j+1}} |x - c_j|\, dx = \sum_{j=0}^{n} \mathcal{E}_j\ ,$$

where

$$\mathcal{E}_j = E\int_{X_j}^{X_{j+1}} |x-c_j|\,dx$$

$$(15) \qquad = \tfrac{1}{2}E\left[(X_{j+1}-c_j)\left|X_{j+1}-c_j\right| - (X_j-c_j)\left|X_j-c_j\right|\right].$$

Now

$$(16) \qquad E\left[(X_j-c_j)\left|X_j-c_j\right|\right] = \int_{c_j}^{1} (y-c_j)^2\, p(y)dy - \int_{0}^{c_j} (y-c_j)^2\, p(y)dy$$

where p(y) is given by (8), and similarly we can get $E\left[(X_{j+1}-c_j)\left|X_{j+1}-c_j\right|\right]$.

Substituting in (15) we easily obtain

$$(17) \qquad \begin{aligned} \mathcal{E}_j = \tfrac{1}{2}\binom{n}{j}\Big[&(n-j)\int_{c_j}^{1}(y-c_j)^2 y^j (1-y)^{n-j-1}\,dy \\ &- (n-j)\int_{0}^{c_j}(y-c_j)^2\, y^j\, (1-y)^{n-j-1}dy \\ &- j\int_{c_j}^{1} (y-c_j)^2\, y^{j-1}(1-y)^{n-j}\,dy \\ &+ j\int_{0}^{c_j} (y-c_j)^2\, y^{j-1}(1-y)^{n-j}\,dy. \end{aligned}$$

This eventually leads to

$$(18) \qquad \mathcal{E}_j = \binom{n}{j}\left[B(j+2,n-j+1) - c_j B(j+1,n-j+1) + 2\sum_{k=0}^{n-j}(-1)^k\binom{n-j}{k}\frac{c_j^{\,j+k+2}}{(j+k+2)\,(j+k+1)}\right]$$

for $j = 0,1,2,\ldots,n$.

Since $R = \sum_{j=0}^{n}\mathcal{E}_j$, and from (15) we see that for each j, $\mathcal{E}_j$ is positive and depends only on j, it is obvious that to minimize R, it is necessary and sufficient to minimize each $\mathcal{E}_j$ separately. We have

$$(19)\quad \frac{\partial \mathcal{E}_j}{\partial c_j} = 2\binom{n}{j}\left[-\tfrac{1}{2}B(j+1,n-j+1) + \sum_{k=0}^{n-j}(-1)^k\binom{n-j}{k}\frac{c_j^{\,j+k+1}}{j+k+1}\right]$$

$$= 2\binom{n}{j}\left[\int_0^{c_j} z^j(1-z)^{n-j}\,dz - \tfrac{1}{2}B(j+1,n-j+1)\right]$$

and

$$(20)\quad \frac{\partial^2 \mathcal{E}_j}{\partial c_j^{\,2}} = 2\binom{n}{j}\, c_j^{\,j}\,(1-c_j)^{n-j}\ .$$

Setting $\frac{\partial \mathcal{E}_j}{\partial c_j} = 0$ and solving we obtain c_j as the median of the Beta distribution with density

$$(21)\quad g(z) = \frac{1}{B(j+1,n-j+1)}\, z^j(1-z)^{n-j}\ ,\ 0 \qquad 1\ ,$$

for $j=0, 1,2,\ldots,n$.

Since (20) shows that $\frac{\partial^2 \mathcal{E}_j}{\partial c_j^{\,2}} > 0$ for $0 < c_j < 1$, it follows that this solution for c_j in fact minimizes $\mathcal{E}_j$ for $j=0,1,\ldots,n$, and hence minimizes R. The minimax invariant procedure is thus to estimate F(x) by

$$\hat{F}(x) = c_j \quad ; \quad X_j \leq x < X_{j+1}$$
$$j=0,\ 1,\ldots,n,$$

where $(X_1,X_2,\ldots,X_n)$ is the ordered sample, X_0 and X_{n+1} stand for $-\infty$ and $+\infty$ respectively, and c_j $(j=0,1,\ldots,n)$ is the median of the Beta distribution with density (21). An alternative way of obtaining this result is given in section 9. It is rather interesting to note that for the loss function discussed in the last Section, c_j was obtained as the mean of the same Beta distribution.

The actual computation of the values of c_j $(j=0,1,\ldots n)$ can be easily carried out, for a given n, with the help of the tables of the incomplete Beta-function [3]. In the notation of the tables

(22) $$I_x(p,q) = \frac{\int_0^x x^{p-1}(1-x)^{q-1}dx}{\int_0^1 x^{p-1}(1-x)^{q-1}dx} .$$

Thus we have to find the value of x such that

(23) $$I_x(j+1, n-j+1) = \tfrac{1}{2}$$

Using the relation

(24) $$I_x(p,q) = 1 - I_{1-x}(q,p) ,$$

it is seen that

(25) $$c_{n-j} = 1 - c_j$$

and thus only about half the total number of c values have to be actually obtained from the tables. The values of c_j $(j = 0,1, \ldots,n)$ for $n = 1,2, \ldots,12$ to two decimal places have been computed and tabulated below.

Table 1

Values of c_j $(j=0,1,\ldots,n)$

for $n=1,2,\ldots.12$

$n \backslash c_j$	c_0	c_1	c_2	c_3	c_4	c_5	c_6	c_7	c_8	c_9	c_{10}	c_{11}	c_{12}
1	.29	.71											
2	.21	.50	.79										
3	.16	.39	.61	.84									
4	.13	.31	.50	.69	.87								
5	.11	.26	.42	.58	.74	.89							
6	.09	.23	.36	.50	.64	.77	.91						
7	.08	.20	.32	.44	.56	.68	.80	.92					
8	.07	.18	.28	.39	.50	.61	.72	.82	.93				
9	.07	.16	.26	.35	.45	.55	.65	.74	.84	.93			
10	.06	.15	.23	.32	.41	.50	.59	.68	.77	.85	.94		
11	.06	.14	.22	.30	.38	.46	.54	.62	.70	.78	.86	.94	
12	.05	.13	.20	.27	.35	.42	.50	.58	.65	.73	.80	.87	.95

5. The loss function $L(F,\hat{F}) = \int_{-\infty}^{\infty} \left[F(x) - \hat{F}(x)\right]^2 / F(x)\left[1-F(x)\right] dF(x)$.

For this loss function, as mentioned before, we get

$$R = E \sum_{j=0}^{n} \int_{X_j}^{X_{j+1}} \frac{(x-c_j)^2}{x(1-x)}\, dx$$

$$(26) \qquad = E\int_0^{X_1} \frac{(x-c_0)^2}{x(1-x)}\, dx + \sum_{j=1}^{n-1} E\Big[-(X_{j+1}-X_j) + c_j^2(\log X_{j+1} - \log X_j) - (1-c_j)^2\left\{\log(1-X_{j+1}) - \log(1-X_j)\right\}\Big] + E\int_{X_n}^{1} \frac{(x-c_n)^2}{x(1-x)}\, dx.$$

For finite risk the integrals in the first and the last terms of the above expression must be finite. The necessary and sufficient condition for this is that $c_0 = 0$ and $c_n = 1$. Our set c_j $(j = 0,1,\ldots,n)$ must then be such that $c_0 = 0$ and $c_n = n$.

With the convention that $0 \times \infty = 0$, we can, therefore, write (26) in the form

$$(27) \qquad R = \sum_{j=0}^{n} E\Big[-(X_{j+1}-X_j) + c_j^2(\log X_{j+1} - \log X_j) - (1-c_j)^2\left\{\log(1-X_{j+1}) - \log(1-X_j)\right\}\Big].$$

The probability density of X_j is given by (8), from which we obtain

$$(28) \qquad E(\log X_j) = j\binom{n}{j}\int_0^1 y^{j-1}(1-y)^{n-j}\log y\, dy$$

and

$$(29) \qquad E(\log(1-X_j)) = j\binom{n}{j}\int_0^1 y^{j-1}(1-y)^{n-j}\log(1-y)\, dy$$

In order to evaluate (28) and (29) we make use of the following lemmas:

Lemma 5.2:

$$(30)\quad \int_0^1 y^{j-1}(1-y)^{n-j}\log y\,dy = \frac{\Gamma(j)\Gamma(n-j+1)}{\Gamma(n+1)}\left[\psi(j)-\psi(n+1)\right] \text{ where } \psi(k)=\Gamma'(k)/\Gamma(k).$$

Proof. Let $f(\alpha)=\int_0^1 y^{\alpha-1}(1-y)^{n-j}\,dy$. The left hand side of (30) is $f'(\alpha)$ evaluated at $\alpha=j$ as can be seen by differentiating under the integral sign. But $f(\alpha) = \Gamma'(\alpha)\ \Gamma(n-j+1)/\Gamma(\alpha+n-j+1)$, and the desired result is obtained by evaluating the logarithmic derivative of $f(\alpha)$ at $\alpha=j$.

Lemma 5.2:

$$(31)\quad \int_0^1 y^{j-1}(1-y)^{n-j}\log(1-y)\,dy = \frac{\Gamma(j)\Gamma(n-j+1)}{\Gamma(n+1)}\left[\psi(n-j+1)-\psi(n+1)\right]$$

where $\psi(k)=\Gamma'(k)/\Gamma(k)$.

Proof. Exactly as in Lemma 5.1, or easily obtained from it by a change of variables.

Utilizing Lemmas 5.1 and 5.2, we obtain

$$(32)\quad E(\log X_j) = \psi(j) - \psi(n+1)$$

and

$$(33)\quad E\log(1-X_j) = \psi(n-j+1) - \psi(n+1),$$

where $\psi(k)=\Gamma'(k)/\Gamma(k)$.

Further, since $\Gamma(k+1) = k\Gamma(k)$, $\Gamma'(k+1) = \Gamma(k) + k\Gamma'(k)$, we see that $\psi(k+1) = \Gamma'(k+1)/\Gamma(k+1) = 1/k + \psi(k)$, and hence the function ψ satisfies the difference equation

$$(34)\quad \psi(k+1) - \psi(k) = 1/k.$$

From (32), (33) and (34) we get

$$(35)\quad E(\log X_{j+1} - \log X_j) = 1/j \quad , \text{ for } j \neq 0$$

and

$$(36)\quad E\left[\log(1-X_{j+1}) - \log(1-X_j)\right] = -1/(n-j) \ , \text{ for } j \neq n.$$

Substituting from (10), (35), and (36) in (27) we obtain at once

$$(37) \qquad R = \frac{2}{n(n+1)} + \sum_{j=1}^{n-1}\left[-\frac{1}{n+1} + \frac{1}{j}c_j^2 + \frac{1}{n-j}(1-c_j)^2\right] = \frac{1}{n} + \sum_{j=1}^{n-1}\frac{n}{j(n-j)}(c_j - \frac{j}{n})^2$$

It is seen from (37) that R is minimized by choosing

$$(38) \qquad c_j = j/n \quad \text{for } j=1,2,\ldots,(n-1).$$

Since $c_0 = 0$ and $c_n = 1$, this expression for c_j holds good also for $j = 0$, and $j = n$. Thus the minimax invariant estimate $\hat{F}$ for the loss function in this Section turns out to be the usual sample cumulative function

$$(39) \qquad \hat{F}(x) = c_j = j/n, \text{ when } X_j \leq x < X_{j+1}, \quad j=0,1,\ldots,n,$$

where $X_1 < X_2 < \ldots . < X_n$ is an ordered sample from the c.d f. F, X_0 and X_{n+1} standing for $-\infty$ and $+\infty$ respectively. The actual value of the risk corresponding to this estimate is $1/n$.

6. The Loss function $L(F,\hat{F}) = \int_{-\infty}^{\infty}\left|F(x) - \hat{F}(x)\right| / F(x)\left[1-F(x)\right] dF(x)$.

In this case we obtain

$$(40) \qquad R = E\sum_{j=0}^{n}\int_{X_j}^{X_{j+1}}\left|x-c_j\right| / x(1-x)dx = \sum_{j=0}^{n}\mathcal{E}_j \quad ,$$

where

$$(41) \quad \mathcal{E}_j = E\int_{X_j}^{X_{j+1}}\left|x-c_j\right| / x(1-x)\ dx$$

As in the last Section, it will be seen that for finite risk the necessary and sufficient condition is that $c_0 = 0$ and $c_n = 1$. For $j \neq 0$, n, we obtain

$$(42)\quad \mathcal{E}_j = E\left[c_j\left|\log c_j - \log X_j\right| - c_j\left|\log c_j - \log X_{j+1}\right| + (1-c_j)\left|\log(1-c_j) - \log(\right.\right.$$

$$\left.\left. -(1-c_j)\left|\log(1-c_j) - \log(1-X_j)\right|\right]\right].$$

The distribution of X_j has probability density $p(y)$ given by (8) and th distribution of X_{j+1} has the probability density

$$(43)\quad q(y) = \frac{1}{B(j+1,n-j)}\, y^j\,(1-y)^{n-j-1}\ ,\ 0 \leq y \leq 1.$$

Using (8) and (43) we can express $\mathcal{E}_j$ in the form

$$(44)\quad \mathcal{E}_j = \binom{n}{j}\left[\int_0^{c_j}\phi(c_j,y)\ dy - \int_{c_j}^{1}\phi(c_j,y)\ dy\right]$$

where

$$(45)\quad \phi(c_j,y) = \left[c_j\log c_j + (1-c_j)\log(1-c_j) - c_j\log y - (1-c_j)\log(1-y)\right]y^{j-1}(1$$

Straightforward integration leads to

$$(46)\quad \int\phi(c_j,y)dy = y^j(1-y)^{n-j}\left[c_j(\log c_j - \log y) + (1-c_j)(\log(1-c_j) - \log(1-\right.$$

$$+ \int(c_j-y)y^{j-1}\ (1-y)^{n-j-1}\ dy\ \ +\ \text{constant}$$

which enables us to obtain $\mathcal{E}_j$ as

$$(47)\quad \mathcal{E}_j = \binom{n}{j}\left[\int_0^{c_j}(c_j-y)y^{j-1}\ (1-y)^{n-j-1}\ dy - \int_{c_j}^{1}(c_j-y)y^{j-1}(1-y)^{n-j-1}d\right.$$

for $j = 1,2,\ldots,\ n-1.$

Since $\mathcal{E}_0$ and $\mathcal{E}_n$ are fixed, and each $\mathcal{E}_j$ is positive and depends only to minimize R it is clearly necessary and sufficient to minimize $\mathcal{E}_j$ j. We see that ution

follows

ence

function

d for

cy (50).

the last

tion.

ositive

$$(52)\quad \mathcal{E}_j = \frac{1}{2s+1}\ E \sum_{k=0}^{2s+1} \binom{2s+1}{k} (-c_j)^{2s+1-k} (X_{j+1}^k - X_j^k)$$

for $j = 0,1,2,\dots,n$.

Substituting from (10) in (52) we obtain

$$(53)\quad \begin{aligned} \mathcal{E}_j &= \frac{1}{n+1}\ c_j^{2s} + \frac{1}{2s+1} \sum_{k=2}^{2s+1} \binom{2s+1}{k} (-c_j)^{2s+1-k} \frac{k(j+1)\dots(j+k-1)}{(n+1)\dots(n+k)} \\ &= \frac{1}{n+1} \left[c_j^{2s} + \sum_{k=2}^{2s+1} \binom{2s}{k-1} (-c_j)^{2s+1-k} \frac{(j+1)\dots(j+k-1)}{(n+1)\dots(n+k)} \right] \end{aligned}$$

For conciseness we introduce the following notation somewhat similar to the binomial and distinguished from it by an asterisk:

$$(54)\quad \left(t - \frac{a+1}{b+1}\right)^{q*} = t^q + \sum_{k=1}^{q} (-1)^k \binom{q}{k} t^{q-k} \prod_{i=1}^{k} \frac{a+i}{b+i}$$

for fixed real a and b and a positive integer q.

It is easily verified that for any positive integer r,

$$(55)\quad \frac{d^r}{dt^r}\left(t - \frac{a+1}{b+1}\right)^{q*} = q(q-1)\dots(q-r+1)\left(t - \frac{a+1}{b+1}\right)^{(q-r)*} \text{ when } r \leq q$$
$$= 0 \text{ when } r > q.$$

Using this notation we can write

$$(56)\quad \mathcal{E}_j = \frac{1}{n+1}\left(c_j - \frac{j+1}{n+2}\right)^{2s*}$$

We have to choose c_j so as to minimize R. Obviously minimizing R is equivalent to minimizing $\mathcal{E}_j$ separately for each j. We obtain

$$(57)\quad \frac{\partial \mathcal{E}_j}{\partial c_j} = \frac{2s}{n+1}\left(c_j - \frac{j+1}{n+2}\right)^{(2s-1)*}, \text{ and}$$

$$(58)\quad \frac{\partial^2 \mathcal{E}_j}{\partial c_j^2} = \frac{2s(2s-1)}{n+1}\left(c_j - \frac{j+1}{n+2}\right)^{(2s-2)*} .$$

Since $\mathcal{E}_j = E\int_{X_j}^{X_{j+1}} (x-c_j)^{2s}\,dx > 0$, it is clear that

(59) $$\frac{\partial^2 \mathcal{E}_j}{\partial c_j^{\,2}} = 2s(2s-1)\ E \int_{X_j}^{X_{j+1}} (x-c_j)^{2s-2} dx > 0.$$

Let $f(c_j) = \frac{\partial \mathcal{E}_j}{\partial c_j}$. It is easily seen that f(0) is negative and f(1) is positive, and since $f'(c_j) > 0$ for all real c_j, $f(c_j)$ is a strictly increasing function of c_j. Hence $f(c_j) = 0$ for one and only one real value of c_j, and this c_j necessarily lies between 0 and 1. Thus we find that $\mathcal{E}_j$, and hence R_2 is minimized by setting $\frac{\partial \mathcal{E}_j}{\partial c_j} = 0$ and solving for c_j the resulting equation

(60) $$(c_j - \frac{j+1}{n+2})^{(r-1)*} = 0$$

This equation has one and only one real root which lies between 0 and 1. The minimax invariant procedure for the loss function of this Section is thus to estimate F(x) by

$$\hat{F}(x) = c_j \quad ; \quad X_j \leq x < X_{j+1}$$
$$j = 0, 1, \ldots n,$$

where X_j ; $j = 0, 1, \ldots, n+1$, have been defined earlier and c_j is the real root of (60). It can further be seen from (60) that the equation remains unchanged if we replace j by n-j and c_j by $1-c_j$. Hence $c_{n-j} = 1-c_j$, and we see that in practice the number of equations to be solved is about half the sample size.

It may be noticed that for r=2, the equation (60) reduces to a linear equation

(61) $$(c_j - \frac{j+1}{n+2})^{1*} = 0$$

which has the unique solution $c_j = \frac{j+1}{n+2}$ as obtained earlier in (12).

8. The loss function $L(F,\hat{F})= \int_{-\infty}^{\infty} \left[F(x)-\hat{F}(x)\right]^{r}/F(x)\left[1-F(x)\right]dF(x)$ where r is any positive even integer.

Let $r=2s$, then

$$R=E\sum_{j=0}^{n}\int_{X_j}^{X_{j+1}}\frac{(x-c_j)^{2s}}{x(1-x)}dx=\sum_{j=0}^{n}\mathcal{E}_j \quad , \tag{62}$$

where

$$\mathcal{E}_j=E\int_{X_j}^{X_{j+1}}\frac{(x-c_j)^{2s}}{x(1-x)}\ dx\ . \tag{63}$$

Since $X_0=0$ and $X_{n+1}=1$, it is clear that in order to obtain finite risk it is necessary and sufficient that $c_0=0$ and $c_n=1$. For $j\neq 0$, n, we can write

$$\mathcal{E}_j=E\left[\sum_{h=0}^{2s-2}\frac{1}{h+1}\ a_h\ (X_{j+1}^{h+1}-X_j^{h+1})+c_j^{2s}(\log X_{j+1}-\log X_j) -(1-c_j)^{2s}\left\{\log\ (1-X_{j+1})-\log\ (1-X_j)\right\}\right] \tag{64}$$

where

$$a_h=-\sum_{i=0}^{2s-2-h}\binom{2s}{i}(-c_j)^{i}\quad ;\quad h=0,1,2,\ldots,(2s-2). \tag{65}$$

Substituting from (10), (35) and (36), we get

$$\mathcal{E}_j=\sum_{h=0}^{2s-2}\frac{(j+h)!n!}{(n+h+1)!j!}\ a_h+\frac{1}{j}\ c_j^{2s}+\frac{1}{n-j}\ (1-c_j)^{2s} \tag{66}$$

and substituting from (65), we can write

$$\mathcal{E}_j=\frac{n!}{j!}\sum_{h=0}^{2s-2}\frac{(j+h)!}{(n+h+1)!}\ \sum_{i=0}^{2s-2-h}\ (-1)^{i+1}\binom{2s}{i}\ c_j^{i}+\frac{1}{j}c_j^{2s}+\frac{1}{n-j}(1-c_j)^{2s}\ . \tag{67}$$

This is a 2sth degree polynomial in c_j. Collecting the coefficients of like powers of c_j we obtain

(68) $$\mathcal{E}_j = \frac{n}{j(n-j)}\, c_j^{2s} - \frac{2s}{n-j}\, c_j^{2s-1} + \sum_{k=0}^{2s-2} \varepsilon_k c_j^k ,$$

where

(69) $$\varepsilon_k = (-1)^{k+1}\binom{2s}{k}\left[\frac{n!}{j!}\sum_{h=0}^{2s-2-k}\frac{(j+h)!}{(n+h+1)!} - \frac{1}{n-j}\right]$$

for $k=0,1,2,\ldots, 2s-2$.

To simplify (68) further, we state and prove the following lemma:

Lemma 8.1. If j and n are positive integers and $j<n$, then

$$\frac{n!}{j!}\sum_{h=0}^{q}\frac{(j+h)!}{(n+h+1)!} = \frac{1}{n-j}\left[1 - \prod_{\alpha=1}^{q-1}\frac{j+\alpha}{n+\alpha}\right] ,$$

Proof. The left hand side

$$= \binom{n}{j}\sum_{h=0}^{q}\frac{(j+h)!\,(n-j)!}{(n+h+1)!}$$

$$= \binom{n}{j}\sum_{h=0}^{q}\int_0^1 x^{j+h}\,(1-x)^{n-j}dx$$

$$= \binom{n}{j}\int_0^1 (x^j - x^{j+q+1})\,(1-x)^{n-j-1}dx$$

= the right hand side, after simplification.

Substituting in (69) from the lemma when $q = 2s-2-k$, we obtain

(70) $$\varepsilon_k = (-1)^k \frac{1}{n-j}\binom{2s}{k}\prod_{\alpha=1}^{2s-1-k}\frac{j+\alpha}{n+\alpha} \quad \text{for } k=0,\ 1,\ 2,\ldots,\ 2s-2,$$

and substituting now in (68) we obtain

(71) $$\mathcal{E}_j = \frac{n}{j(n-j)}\left[c_j^{2s} + \sum_{k=0}^{2s-1} \binom{2s}{k} (-c_j)^k \prod_{\alpha=0}^{2s-1-k} \frac{j+\alpha}{n+\alpha}\right] = \frac{n}{j(n-j)} \quad (c_j - \frac{j}{n})^{2s*}$$

in the notation introduced in (54).

Now with the same reasoning as in the last section it will be seen that $\mathcal{E}_j$ and hence R is minimized by setting $\frac{\partial \mathcal{E}_j}{\partial c_j} = 0$ and solving for c_j the resulting equation

(72) $$(c_j - j/n)^{(r-1)*} = 0.$$

This equation has one and only one real root which lies between 0 and 1. Since for $j=0$, (72) reduces to $c_0^{r-1} = 0$ giving $c_0 = 0$ as the only real root, and for $j=n$, it reduces to $(c_n - 1)^{r-1} = 0$, giving $c_n = 1$ as the only real root , it follows that we can say that the minimax invariant proceedure for the loss function of this Section is to estimate F(x) by

$$\hat{F}(x) = c_j \quad ; \quad X_j \leq x < X_{j+1}$$
$$j = 0, 1, \ldots, n,$$

where X_j ; $j=0, 1, \ldots, n+1$ have been defined earlier and c_j is the real root of (72). Again the number of equations to be solved in practice will be about half the sample size since it can be easily seen that (72) remains unchanged by replacing j by n-j and c_j by $1-c_j$, so that $c_{n-j} = 1-c_j$.

For r=2, the equations (72) reduces to $c_j - j/n = 0$ giving $c_j = j/n$ for all j.

9. The loss function $L(F,\hat{F}) = \int_{-\infty}^{\infty} |F(x)-\hat{F}(x)|^r dF(x)$, where r is any positive integer.

In this case

(73) $$R = E \sum_{j=0}^{n} \int_{X_j}^{X_{j+1}} |x - c_j|^r \, dx = \sum_{j=0}^{n} \mathcal{E}_j ,$$

where

$$(74) \quad \mathcal{E}_j = \frac{1}{r+1} \; E\left[(X_{j+1} - c_j)\left|X_{j+1} - c_j\right|^r - (X_j - c_j)\left|X_j - c_j\right|^r\right].$$

Using (8) we obtain

$$(75) \quad E\left[(X_j - c_j)\left|X_j - c_j\right|^r\right] = j\binom{n}{j}\left[\int_{c_j}^{1} (y-c_j)^{r+1} \, y^{j-1} \, (1-y)^{n-j} dy - \int_0^{c_j} (c_j - y)^{r+1} \, y^{j-1} \, (1-y)^{n-j} \, dy\right],$$

and similarly,

$$(76) \quad E\left[(X_{j+1} - c_j)\left|X_{j+1} - c_j\right|^r\right] = (n-j) \binom{n}{j}\left[\int_{c_j}^{1} (y-c_j)^{r+1} \, y^j (1-y)^{n-j-1} \, dy - \int_0^{c_j} (c_j - y)^{r+1} \, y^j (1-y)^{n-j-1} dy\right].$$

From (75) and (76) we obtain

$$(77) \quad \mathcal{E}_j = \frac{1}{r+1} \binom{n}{j}\left[\int_{c_j}^{1} (y-c_j)^{r+1} \, y^{j-1} \, (1-y)^{n-j-1} \, (ny-j)dy + (-1)^r \int_0^{c_j} (y-c_j)^r \quad y^{j-1} \, (1-y)^{n-j-1} (ny-j)dy\right]$$

Again it is obvious that to minimize R is equivalent to minimizing $\mathcal{E}_j$ for each j. Further we see that the conditions for differentiation with respect to c_j under the integral sign in (77) are satisfied, and we obtain

$$(78) \quad \frac{\partial \mathcal{E}_j}{\partial c_j} = -\binom{n}{j}\left[\int_{c_j}^{1} (y-c_j)^r \, y^{j-1} \, (1-y)^{n-j-1} (ny-j)dy + (-1)^r \int_0^{c_j} (y-c_j)^r \, y^{j-1} \, (1-y)^{n-j-1} (ny-j)dy\right]$$

and for $r \geq 2$,

$$\frac{\partial^2 \mathcal{E}_j}{\partial c_j^{\,2}} = r\binom{n}{j}\left[\int_{c_j}^{1}(y-c_j)^{r-1}y^{j-1}(1-y)^{n-j-1}(ny-j)dy + (-1)^r\int_0^{c_j}(y-c_j)^{r-1}y^{j-1}(1-y)^{n-j-1}(ny-j)dy\right] \tag{79}$$

$$= r(r-1)\,E\int_{x_j}^{x_{j+1}}\left|x-c_j\right|^{r-2}dx > 0$$

As in fact on 4, if we define a function ϕ, $\phi(c_j) = \frac{\partial \mathcal{E}_j}{\partial c_j}$, we see it [illegible] [illegible] for all [illegible] that

$$\phi(0) = r\frac{n!(r+j-1)!}{j!(r+n)!} < 0$$

and

$$\phi(1) = r\frac{n!(r+n-j-1)!}{(n-j)!(r+n)!} > 0$$

and further $\phi'(c_j) > 0$ for all real c_j. Thus ϕ is a strictly increasing function of c_j, assumes the value zero for one and only one real value of c_j, and this value of c_j necessarily lies between zero and one. Thus we find that $\mathcal{E}_j$ and hence R is minimized by setting $\frac{\partial \mathcal{E}_j}{\partial c_j} = 0$ and solving for c_j the resulting equation

$$\int_{c_j}^{1}(y-c_j)^r y^{j-1}(1-y)^{n-j-1}(ny-j)dy + (-1)^r\int_0^{c_j}(y-c_j)^r y^{j-1}(1-y)^{n-j-1}(ny-j)dy = 0 \tag{80}$$

Thus the problem reduces to that of solving the above equation for $j = 0, 1, \ldots, n$. The general solution of (80) giving c_j explicitly in terms of j, n, and r does not seem to be possible. We shall, however, simplify the equation so that it should not be too difficult to obtain the solution in any given case. It can, however, be proved from (80) that $c_{n-j} = 1-c_j$ so that the number of equations to be solved in practice will be about half the sample size.

We can write (80) as

$$(81)\quad \int_0^1 (y-c_j)^r y^{j-1}(1-y)^{n-j-1}(ny-j)dy = \left\{1-(-1)^r\right\}\int_0^{c_j} (y-c_j)^r y^{j-1}(1-y)^{n-j-1}(ny-j)dy.$$

It would be seen that the right hand side of this equation would be of importance only when r is odd, for when r is even, it reduces to zero.

The left hand side of equation (81) can be expressed as

$$(82)\qquad \sum_{k=0}^{r} k\binom{r}{k}(-c_j)^{r-k}B(j+k,\ n-j+1)\quad ,$$

which indicates that the coefficient of c_j^r is zero for $k=0$. We can utilize the fact that $\binom{r}{k}k = r\binom{r-1}{k-1}$ and reduce it further to the form

$$(83)\qquad rB(j, n-j+1)\sum_{k=1}^{r}\binom{r-1}{k-1}(-c_j)^{r-k}\frac{j(j+1)\cdots(j+k-1)}{(n+1)(n+2)\cdots(n+k)}\quad ,$$

which by making use of the notation introduced in (54) can be written as

$$(84)\qquad (-1)^{r-1}rB(j+1,\ n-j+1)\left(c_j - \frac{j+1}{n+1}\right)^{(r-1)}.$$

As mentioned before, when r is even, the right hand side of equation (81) reduces to zero and cancelling out the non-zero coefficient $(-1)^{r-1}rB(j+1, n-j+1)$ from the left hand side as expressed by (84) we obtain c_j as a root of the same equation (60) obtained earlier by a different method.

The right hand side of the equation (81), except for the factor $\left[1-(-1)^r\right]$, can be written as

$$(85)\qquad \sum_{k=0}^{r}\binom{r}{k}(-c_j)^{r-k}\int_0^{c_j}\sum_{s=0}^{n-j}(-1)^{s-1}(j+s)\binom{n-j}{s}y^{k+j+s-1}\,dy,$$

and by making use of the relation

$$(86)\qquad \sum_{k=0}^{r}(-1)^k\binom{r}{k}\frac{1}{k+t} = B(t,\ r+1)$$

it can be reduced to

$$(87)\qquad (-1)^{r-1} r \sum_{s=0}^{n-j} (-1)^{s} \binom{n-j}{s}\ B\ (r,\ j+s+1) c_j^{r+j+s} \quad .$$

Using (84) and (85) we can, thus, write the equation (81) as

$$(88)\qquad B(j+1,\ n-j+1)\left(c_j - \frac{j+1}{n+2}\right)^{(r-1)*} = \left[1-(-1)^{r}\right] \sum_{s=0}^{n-j} (-1)^{s}\binom{n-j}{s} B(r,\ j+s+1) c_j^{r+j+s} \ .$$

This equation is to be solved for c_j to get a minimax invariant procedure for estimating F when the loss function is given by (1). When r is even, the factor $1-(-1)^r = 0$ and we get an equation of degree (r-1). When r is odd, the factor $1-(-1)^r = 2$ and the equation reduces to

$$(89)\qquad \sum_{s=0}^{n-j} (-1)^{s}\binom{n-j}{s} B(r,\ j+s+1) c_j^{r+j+s} - \tfrac{1}{2} B(j+1,\ n-j+1)\left(c_j - \frac{j+1}{n+2}\right)^{(r-1)*} = 0$$

which is an equation of degree (n+r). In either case there is one and only one real root which lies between 0 and 1 and the set of such roots for j = 0, 1,...,n minimizes R.

An alternative way of expressing the right hand side of (81) is to rewrite (87) in the following form:

$$(-1)^{r-1}\ r! \sum_{s=0}^{n-j} (-1)^{s}\binom{n-j}{s} \frac{c_j^{r+j+s}}{(j+s+1)(j+s+2)\ldots(j+s+r)}$$

$$(90)\qquad = (-1)^{r-1}\ r! \sum_{s=0}^{n-j} (-1)^{s}\binom{n-j}{s} \int_0^{c_j}\int_0^{z_r}\cdots\int_0^{z_2} z_1^{j+s}\, dz_1 \ldots dz_r$$

$$= (-1)^{r-1}\ r! \int_0^{c_j}\int_0^{z_r}\cdots\int_0^{z_2} z_1^{j}(1-z_1)^{n-j} dz_1 \ldots dz_r \ .$$

The equation (88) can, therefore, also be expressed as

(91) $$B(j+1,\ n-j+1)\left(c_j-\frac{j+1}{n+2}\right)^{(r-1)*}=\left[1-(-1)^r\right](r-1)!\int_0^{c_j}\int_0^{z_r}\cdots\int_0^{z_2} z_1^j(1-z_1)^{n-j}dz_1\ldots dz_r.$$

From this form, it is easily seen that for $r=1$, the equation reduces to

(92) $$B(j+1,\ n-j+1)=2\int_0^{c_j} z^j(1-z)^{n-j}\,dz$$

which shows that c_j is the median of the Beta distribution with density

$$g(z)=\frac{1}{B(j+1,\ n-j+1)}\,z^j(1-z)^{n-j}\quad,\ 0\leq z\leq 1$$

for $j=0,\ 1,\ldots,\ n,$

as obtained earlier in Section 4.

I would like to acknowledge some very helpful discussions with Professors Z. W. Birnbaum and H. Rubin during the preparation of this paper.

REFERENCES

[1] Z. W. Birnbaum, "Distribution-Free Tests of Fit for Continuous Distribution Functions," Ann. Math. Stat., Vol. 2[illegible] [illegible]

[2] Z. W. Birnbaum and [illegible], "[illegible]," [illegible]

[3] [illegible], "[illegible]," [illegible]

Institute for Numerical Analysis
405 Hilgard Avenue
Los Angeles 24, California 1

Chief, Statistical Engineering
Laboratory
National Bureau of Standards
Washington 25, D. C. [illegible]

Rand Corporation
1500 Fourth Street
Santa Monica, California [illegible]

[illegible]

[illegible]

Professor [illegible]
[illegible]
[illegible] Johns [illegible] University
Baltimore 18, Maryland

Professor [illegible]
Department of [illegible]
[illegible] University
[illegible] 1

Professor Herbert Solomon
Teachers College
Columbia University
New York, New York 1

Professor W. Allen Wallis
Committee on Statistics
University of Chicago
Chicago 37, Illinois 1

Professor J. Wolfowitz
Department of Mathematics
Cornell University
Ithaca, New York 1

Department of Mathematical Statistics
University of North Carolina
Chapel Hill, North Carolina 2

Professor J. Neyman
Statistical Laboratory
University of California
Berkeley, California 1

[illegible]

Institute for Numerical Analysis
405 Hilgard Avenue
Los Angeles 24, California 2

Chief, Statistical Engineering
Laboratory
National Bureau of Standards
Washington 25, D. C. 1

RAND Corporation
1500 Fourth Street
Santa Monica, California 1

Applied Mathematics and
Statistics Laboratory
Stanford University
Stanford, California 3

Professor Carl B. Allendoerfer
Department of Mathematics
University of Washington
Seattle 5, Washington 1

Professor W. G. Cochran
Department of Biostatistics
The Johns Hopkins University
Baltimore S., Maryland 1

Professor Benjamin Epstein
Department of Mathematics
Wayne University
Detroit 1, Michigan 1

Professor Herbert Solomon
Teachers College
Columbia University
New York, New York 1

Professor W. Allen Wallis
Committee on Statistics
University of Chicago
Chicago 37, Illinois 1

Professor J. Wolfowitz
Department of Mathematics
Cornell University
Ithaca, New York 1

Department of Mathematical Statistics
University of North Carolina
Chapel Hill, North Carolina 2

Professor J. Neyman
Statistical Laboratory
University of California
Berkeley, California 2

Professor S. S. Wilks
Department of Mathematics
Princeton, New Jersey 1